Nature Communications

https://doi.org/10.1038/s41467-026-70421-3

Article in Press

Synthetic data-driven deep learning for label-free autonomous atomic force microscopy

Received: 26 September 2025

Accepted: 26 February 2026

Cite this article as: Millan-Solsona, R., Checa, M., Brown, S.R. *et al.* Synthetic data-driven deep learning for label-free autonomous atomic force microscopy. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-70421-3

Ruben Millan-Solsona, Marti Checa, Spenser R. Brown, Amber N. Bible, Bernadeta Srijanto, Laura Wiggins, Sita Sirisha Madugula, Alice L. B. Pyne, Jennifer L. Morrell-Falvey, Scott Retterer, Rama K. Vasudevan & Liam Collins

Synthetic Data-Driven Deep Learning for Label-free Autonomous Atomic Force Microscopy

Ruben Millan-Solsona,[1] Marti Checa,[1] Spenser R. Brown,[2] Amber N. Bible,[2] Bernadeta Srijanto[1], Laura Wiggins,[3] Sita Sirisha Madugula,[1] Alice L. B. Pyne,[3] Jennifer L. Morrell-Falvey,[2] Scott Retterer,[1,2] Rama K. Vasudevan,[1] Liam Collins.[1*]*

[1]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA.

[2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA.

[3] School of Chemical, Materials and Biological Engineering, University of Sheffield, Sheffield, UK

*Correspondence to: solsonarm@ornl.gov & collinslf@ornl.gov

Funding: U.S. Department of Energy, Office of Science FWP ERKCZ64, Structure Guided Design of Materials to Optimize the Abiotic-Biotic Material Interface

Abstract: Atomic force microscopy (AFM) is a widely used tool for nanoscale characterization across materials science, energy research, and biology. However, its adoption in high-throughput materials discovery and statistically driven studies remains limited by a strong dependence on expert operator input and by the scarcity of annotated experimental AFM datasets needed to enable data-driven automation. Here, we introduce SimuScan, a synthetic-data–driven framework that enables reliable AFM feature identification, segmentation, and targeted imaging without requiring large manually labeled experimental datasets. SimuScan generates tunable, high-fidelity synthetic AFM images of defined morphologies while incorporating realistic experimental artifacts, including tip–sample convolution, noise, flattening distortions, and surface debris. These datasets are shown to support scalable, label-free training of modern deep learning models for AFM analysis. When integrated into data-driven AFM workflows, SimuScan-trained models can locate and analyze nanoscale structures across large datasets and guide targeted follow-up imaging. We validate this approach on nanostructured surfaces, DNA assemblies, and bacterial cells, demonstrating robust generalization across diverse sample types with minimal operator intervention.

More broadly, this work establishes a general strategy for generating explicitly conditioned, task-relevant synthetic data to improve the reliability of downstream models in autonomous microscopy.

1. Introduction

Atomic force microscopy (AFM)[1] provides topographical and functional information with nanometer resolution[2] and has become an indispensable tool across materials science,[2] energy research,[3] and molecular and cellular biology.[4-8] However, wider adoption of AFM remains limited by a strong dependence on expert operator input, slow imaging speeds, and restricted fields of view.[9,10] These constraints severely limit AFM throughput, making large-area exploratory surveys impractical, especially for rare or heterogeneous features, and preventing the collection of statistically meaningful datasets required for high-throughput, discovery-oriented AFM studies. In contrast, optical and electron microscopies can generate large-area surveys in a single acquisition, highlighting AFM's challenges in scalable discovery and quantification. Compounding these issues, image interpretation and quantitative analysis are often time consuming and typically performed offline, further delaying experimental insight.

Recent advances in machine learning (ML) and artificial intelligence (AI) offer a promising route toward more automated AFM workflows, with the potential to accelerate analysis, automate feature recognition, and enable data-driven decision-making during experiments, key to boost scientific discovery.[11] Parallel progress in AFM hardware automation—including motorized stages, programmable interfaces, and automated routines for tip approach and feedback control[12-19]—has laid the groundwork for integrating AI into AFM operation. Early studies have demonstrated the use of ML for scan optimization,[20] parameter tuning,[21-24] and data interpretation, [25-27] while more recent efforts have begun to address higher-level tasks such as autonomous site selection[28-30] and adaptive probing.[31] Collectively, these developments point toward a shift from operator-driven AFM workflows to more adaptive and autonomous experimentation.

Despite this progress, broader deployment of AI-driven AFM remains hindered by a fundamental bottleneck: the scarcity of high-quality, labeled AFM datasets suitable for supervised learning. Unlike optical imaging, where large public datasets (e.g. Imagenet[32]) underpin modern foundation models and transfer-learning approaches, AFM data are comparatively sparse, highly heterogeneous, and strongly shaped by convolution artifacts, noise, drift, and imaging conditions. As a result, existing pretrained foundation models such as Meta's Segment Anything Model (SAM)[33] transfer poorly to AFM data, requiring task-specific retraining with expert-labeled AFM images, which is time intensive,[34] and limits dataset scale and diversity. This bottleneck has slowed the application of deep learning to core AFM tasks such as segmentation, classification, and adaptive site selection.

Although synthetic AFM data have not been widely used for training machine-learning models, they have long supported AFM research by enabling controlled studies of systematic errors, algorithm

benchmarking, and data interpretation.[35] Prior work has used simulated AFM images to investigate tip–sample convolution effects,[36] fractal surfaces,[37] nanoparticle size distributions,[38] and to benchmark denoising[39] and double-tip correction algorithms.[40] Specialized tools[41] have also generated simulated AFM images of biomolecular structures to support structure interpretation and model-based reconstruction.[42,43] At the atomic scale, synthetic AFM data have also enabled deep learning approaches for automated structure discovery[44] and molecular identification. [45] However, most existing synthetic AFM datasets remain tailored to specific samples, tip models, or imaging conditions and are not designed to generate large, diverse, and fully labeled datasets suitable for training general-purpose deep learning models for AI-driven AFM purposes.

To address this gap, we introduce SimuScan, a synthetic-data–driven framework for AFM that enables segmentation, feature identification, and targeted imaging without requiring large manually labeled experimental datasets. SimuScan generates realistic AFM images by combining flexible geometry definitions, including CAD-defined and example-based geometries derived from existing AFM data, with physics-informed modeling of essential AFM-specific artifacts such as object distortions, surface contamination, tip convolution, feedback artifacts, instrumental noise, and flattening distortions. Each synthetic image is produced together with its corresponding ground-truth mask, enabling fully labeled training data without expert manual annotation.

We demonstrate the applicability of SimuScan across a diverse range of nanostructured surfaces, DNA assemblies, and bacterial cells, showing that models trained exclusively on synthetic data generalize effectively to experimental AFM images. Across multiple deep learning segmentation architectures, SimuScan enables reliable feature recognition and prioritization without additional manual labeling, establishing a workflow in which objects of interest are automatically identified based solely on their morphological characteristics. Finally, when integrated with automated tip engagement and stage navigation, SimuScan supports semi-autonomous AFM operation that spans large-area exploratory surveys and targeted high-resolution imaging. More broadly, this work shows how domain-specific synthetic datasets can remove key bottlenecks in machine-learning–enabled experimentation, establishing a scalable path toward intelligent, high-throughput AFM that extends advanced nanoscale characterization beyond expert operators.

2. Results and Discussion

2.1. SimuScan: a generator for labeled realistic synthetic AFM datasets

To overcome the scarcity of training data for AFM segmentation models, SimuScan generates labeled synthetic datasets from arbitrary nanoscale morphologies coupled with realistic imaging artifacts, including tip convolution, flattening distortions, and noise, to reproduce experimental AFM conditions (Figure 1). These images are automatically converted into COCO (Common Objects in Context)-formatted datasets with corresponding masks, making them directly compatible with modern deep learning pipelines. Within the SimuScan workflow, a range of segmentation architectures, including semantic, object detection, and instance segmentation models, can be trained and integrated into automated AFM operation, enabling real-time feature recognition, analysis, and feedback. In the sections that follow, we describe the synthetic AFM data generation framework, demonstrate the realism of the generated data, evaluate models trained exclusively on SimuScan datasets, and conclude by showcasing its integration into a semi-autonomous AFM imaging workflow.

Figure 2 illustrates the SimuScan data generation pipeline, which consists of three main stages: morphology definition, substrate integration, and AFM-specific forward modeling. The process begins with morphology definition, where object shapes are specified using one of three geometry definition modes. SimuScan supports multiple complementary pathways (Figure. 2a) for generating basic sample/object shapes including:

(i) Parameterized model.

(ii) CAD Geometry.

(iii) Experimental (self-Seeding) model.

Together these enable synthetic AFM data generation across a wide range of nanoscale systems with increasing complexity. In Parameterized model, simple shapes can be defined using parameterized, two-dimensional mask projections composed of straight segments and spline curves controlled by reference points. In CAD Geometry and Experimental Geometry, object morphology is provided as a three-dimensional STL (stereolithography) mesh, either imported directly from an externally designed CAD model or derived from an experimentally acquired AFM image by converting measured surface topography into STL format. In the latter case, representative experimental features serve as self-seeds, enabling AFM-to-AFM transfer learning through synthetic dataset generation.

Across all geometry definition modes (See Figure S1 for more details), SimuScan introduces controlled shape variability through stochastic perturbations and spatially correlated deformation fields, including sinusoidal and Gaussian vector fields that impose bending, warping, or curvature with adjustable amplitude. In addition, a range of surface textures and complex features are incorporated to emulate realistic nanoscale heterogeneity commonly observed in experimental AFM images (See Figure

S2). These transformations preserve the underlying morphological identity while introducing realistic variability representative of experimentally acquired AFM datasets.

In the next stage, the synthetic objects are combined with a simulated substrate to form the ground-truth surface geometry (Figure 2b). The substrate can include features such as steps, periodic patterns (e.g., pillars, holes, or gratings), surface roughness, contamination, and random debris, all defined through fully controllable parameters. This enables systematic variation of geometry, periodicity, height, spacing, and contamination density (see Figure S3), emulating substrate features and surface imperfections commonly encountered in experimental AFM measurements. At this point in the workflow, the combined object–substrate system represents an idealized surface morphology prior to AFM-specific distortions.

Finally, AFM imaging artifacts are introduced through a sequence of forward-modeling layers to approximate realistic AFM measurements (Figure 2c; see Section S4 in the Supporting Information). These include tip–sample convolution to model resolution loss due to finite probe geometry, with explicit control over tip radius and cone angle, as well as the ability to simulate single- or double-tip configurations to capture common probe-related artifacts (see Figure S5). Feedback-related effects are added to account for controller-induced distortions such as line drift, discontinuities, and adhesion-related artifacts, followed by electronic noise representative of scanner and signal-path fluctuations. Finally, post-processing effects, including line-flattening errors, are applied to replicate distortions commonly observed in experimental AFM datasets. Although flattening is not an intrinsic part of AFM data acquisition, it is routinely applied to correct drift, scanner tilt, and contrast variations and can introduce secondary artifacts such as residual line features or local distortions. SimuScan has the ability to incorporate flattening-induced artifacts to improve dataset realism and enhance model robustness to post-processing effects encountered in practice. Like all the parameters described, users retain control over whether flattening is applied and over the type and magnitude of the correction; in this workflow, flattening is applied prior to object detection to reflect common automated AFM practice. Because both object morphology and imaging artifacts are explicitly defined within the forward model, SimuScan produces synthetic AFM images with exact ground-truth annotations.

Figures 2d,e shows an example of synthetic AFM image (Figure 2d) and their corresponding labeled map (Figure 2e) for a nanostructured surface, where geometric shaped objects of different height and size are randomly placed and oriented, and realistic artifacts such as noise and line-flattening errors are incorporated. Figure 3 provides additional examples of synthetic datasets, including regular geometric shapes (Figure 3a) and complex biological morphologies such as bacterial cells (Figure 3b), highlighting the flexibility of the framework to capture both ordered and irregular nanoscale structures. Extending this

versatility to microbial systems, SimuScan can also simulate bacterial colonies by first inserting an individual bacterium at a random position and then iteratively adding neighboring cells with controlled overlap. During this process, parameters such as height, eccentricity, orientation, and surface texture are varied, resulting in colony-like assemblies where each bacterium is unique yet collectively organized into realistic clusters that closely resemble experimental observations. The resulting datasets support pre-training and fine-tuning of deep neural networks for segmentation and classification tasks, circumventing the need for large labeled experimental datasets.

While SimuScan captures a broad range of AFM-specific artifacts relevant for realistic synthetic data generation, this first implementation makes deliberate scope choices leaving clear avenues for future extension. At present, the framework focuses on the topographic channel, which is broadly applicable across common AFM modes (e.g., dynamic, static, PeakForce Tapping, and wave-based operation), while additional contrast channels such as phase, amplitude, or spectroscopic signals are not explicitly modeled. Long-timescale thermal and mechanical drift is also not simulated, reflecting the prevalence of closed-loop AFM systems and standard post-processing procedures that mitigate these effects in many experiments. In addition, while SimuScan includes artifacts such as double-tip artifacts, it but does not yet capture intermittent or time-dependent probe contamination. Extending the framework to incorporate additional signal channels, drift dynamics, and evolving probe conditions represents a natural and modular path forward, and would further expand the realism, versatility, and applicability of SimuScan across a wider range of AFM modalities and experimental conditions.

2.2. Performance of Models Trained Exclusively on SimuScan Data

To assess whether SimuScan-generated data can support effective model training without additional manual labeling, we evaluated three widely used segmentation frameworks: YOLOv8, U-Net, and Mask R-CNN. All models were trained on more than 5,000 synthetic images, with representative outputs shown in Supplementary Figure S6. The frameworks were evaluated on experimental AFM images containing fabricated geometric nanostructures (Rectangles, Hexagons, Circles) and microbial morphologies, including rod-shaped *Pseudomonas aeruginosa* PAO1 and spherical *Staphylococcus aureus* NCTC 8325 cells.

Performance was assessed using average precision at 0.5 IoU (AP@0.5) for object-level detection and intersection-over-union (IoU) for pixel-wise segmentation fidelity (Table 1). In addition, global object-level precision, recall, and F1-score were computed by aggregating true positives, false positives, and false negatives across all annotated experimental images, providing an intuitive measure of detection

reliability and miss rate on real AFM data. YOLOv8 (≈11M parameters, <10 ms inference) was evaluated using AP@0.5, U-Net (≈7M parameters, ~20 ms) using IoU, and Mask R-CNN (>40M parameters, ~100 ms) using both AP@0.5 and IoU to enable object- and pixel-level benchmarking. Together, these metrics capture detection accuracy, segmentation quality, and sim-to-real transfer performance.

Model performance was evaluated using an independently prepared ground-truth dataset consisting exclusively of manually annotated experimental AFM images; no SimuScan-generated images were included in the evaluation set. Although only 25 images per sample type were annotated, each AFM image contained many repeated instances, yielding 756 geometric objects and 1,124 bacterial instances. Standard data augmentation was applied to introduce controlled variability in orientation, scale, and contrast while preserving morphology, providing a statistically meaningful benchmark for evaluating segmentation performance.

We intentionally did not pursue extensive training or fine-tuning on large numbers of manually annotated experimental AFM images, as the objective of this study was not to maximize task-specific performance through annotation-intensive workflows. In prior AFM segmentation studies using YOLO- or U-Net–based architectures, performance is often improved by training on hundreds of manually labeled experimental images—a process that typically requires days to weeks of expert effort due to careful feature delineation, instance separation, and artifact rejection. In contrast, the central aim of this work is to evaluate whether SimuScan-generated synthetic data can substantially reduce or eliminate this manual labeling burden while still enabling effective model training. In support of this, qualitative inspection of predictions from the manually trained model further confirmed that the level of performance achieved was sufficient for the objectives of the present study, namely assessing the Simuscan data generalization rather than maximizing any task-specific accuracy.

YOLOv8 achieved the highest object-level detection performance, with AP@0.5 values exceeding 0.96 for all geometric shapes and remaining strong for bacilli (0.82) and cocci (0.91). U-Net demonstrated the strongest pixel-level segmentation accuracy, with IoU values approaching 0.90 for nanostructure features and remaining robust across bacterial morphologies (0.80 for Bacilli, 0.71 for Cocci). Mask R-CNN performed competitively for circular and hexagonal features but showed but reduced performance for rectangles and microbial classes. reflecting the increased complexity and conservativeness of its instance-segmentation framework.

These trends are further reflected in the global object-level metrics, where YOLOv8 exhibits the highest recall across both nanostructures (0.98) and bacteria (0.93), indicating a very low rate missed object rate. U-Net displays a balanced precision–recall performance, while Mask R-CNN achieves higher

precision at the expense of recall, consistent with its emphasis on high-confidence instance delineation. Together, these results highlight the complementary strengths of the evaluated models: YOLOv8 is well suited for rapid, object-level detection; U-Net is preferred for accurate boundary segmentation; and Mask R-CNN is advantageous when instance-level masks are required despite higher computational cost. Importantly, the consistently high global F1-scores across both nanostructured and datasets provide quantitative evidence of robust sim-to-real generalization, demonstrating that models trained exclusively on SimuScan-generated data transfer effectively to experimentally acquired AFM images (example outputs are shown in subsequent sections and in Supplementary Information).

To demonstrate the broad applicability of SimuScan, we evaluated the framework across representative case studies spanning simple biological systems to complex engineered nanostructures. As an initial benchmark, a YOLOv8 model trained exclusively on parameterized SimuScan geometries was used to distinguish linear and circular DNA in pre-collected AFM images (Fig. S7). Despite having no exposure to experimental AFM data during training, the model reliably detected DNA molecules and correctly classified their topology without manual annotation or task-specific experimental fine-tuning. This result indicates that SimuScan captures the key morphological features governing DNA conformation, including contour continuity, curvature, and loop closure.[46] Increasing the task complexity, we applied the same parameterized workflow to bacterial cells deposited on a structured substrate patterned with circular nanopillars (Fig. S8). Because SimuScan can be tuned to explicitly incorporate substrate topography during synthetic data generation, models trained on these datasets robustly detected biological cells even on non-planar, topographically heterogeneous surfaces, demonstrating generalization beyond idealized flat geometries.

Beyond parameterized shapes, SimuScan supports two complementary non-parameterized geometry workflows: externally defined CAD geometries and self-seeded experimental geometries. In both cases, object morphology is provided as a CAD/STL representation, enabling large, fully labeled synthetic datasets to be generated without requiring analytical shape models.

In the first case, geometry is defined externally using CAD-designed objects. We demonstrate this capability using complex engineered nanoarchitectures in the form of leaf-like structures generated directly from STL files used in their original design (Fig. S9). These geometries can be incorporated into SimuScan without modification, enabling realistic synthetic AFM datasets for nanofabricated structures that are difficult to capture using parameterized models. Notably, the leaf-like nanoarchitectures were intentionally designed to exhibit subtle but systematic variations and incomplete geometries that are

visually similar, providing a stringent test in which SimuScan-trained models successfully distinguished between closely related structural variants.

In the second case, geometry is derived directly from experimental AFM data using a self-seeded workflow. Representative structures are extracted from experimentally acquired AFM images and converted into CAD/STL geometries, which preserve the true experimental morphology and serve as reference shapes for synthetic data generation. We demonstrate this approach using triangular DNA origami structures. Starting from cropped features extracted from an experimental AFM image (Fig. 4a), representative shapes were converted into STL files and used to generate realistic synthetic training datasets. Models trained on these datasets generalized successfully to experimental AFM images, enabling both object detection and instance segmentation with strong agreement to manual annotations (Fig. 4b).

Across all evaluated sample types, including parameterized models, externally defined CAD geometries, and self-seeded experimental geometries, SimuScan-trained models exhibit consistently strong detection and segmentation performance, following the quantitative trends reported in Table 1. As expected for any cross-domain segmentation task, failure modes arise when experimental conditions fall outside the distribution represented in the synthetic training data, including severe tip wear, drift, corrupted scans, unanticipated feature geometries, or large mismatches in object density, all of which can lead to false detections or missed objects (Figure S10a,b). Performance is also sensitive to surface coverage, with extremely sparse or highly crowded scenes producing missed detections or object merging, as shown for triangular DNA origami in Figure S10c,d. Importantly, these effects are not fundamental limitations but reflect mismatches between the experimental regime and the synthetic training distribution. While these refinements are not the focus of the present work, which instead emphasizes the generality and scalability of SimuScan as a rapid, domain-specific starting point for AFM segmentation, SimuScan is explicitly designed to assist in tuning of the synthetic data generator and model configuration to adapt performance for high-precision or specialized AFM workflows, as illustrated in Figure S10.

2.3. Semi-Autonomous AFM Workflow

Building on this foundation, we integrated SimuScan into a real-time semi-autonomous AFM workflow in which models trained on synthetic data guide feature identification and adaptive scanning during experiments. As shown in Figure 5a and Supplementary Movies 1 and 2, the workflow begins with a low-resolution (LR) overview scan, followed by real-time segmentation to detect and classify individual features. A YOLOv8 model trained exclusively on synthetic AFM data achieved >90% classification

accuracy on overview scans (128 × 128 pixels, ~195 nm per pixel), with resolution-dependent performance quantified in Supplementary Figures S11 and S12.

Detected candidate regions then trigger automated high-resolution (HR) zoom-in scans to verify object presence and morphology. After all targets in the current field of view are evaluated, the system repositions the probe and repeats the cycle of scanning, detection, and validation. Stage automation extends this process beyond the intrinsic ~100 µm AFM scan range,{Millan-Solsona, 2025 #44} enabling tiling across large sample areas and revisiting of features identified by segmentation. User-defined criteria, such as the number of detected objects, surveyed area, confidence score or statistical measures of spatial or morphological variability, are used to determine when further imaging, refinement, or repositioning is required, allowing the experiment to adapt dynamically to the evolving dataset.

This closed-loop workflow enables precise feature identification over large surfaces with minimal human intervention. The approach was validated on both biological and engineered samples, including bacterial cells (Figure 5b) and nanostructured surfaces (Figure 5c), demonstrating robust performance across diverse systems. By directing high-resolution imaging only to regions containing relevant objects, acquisition time is reduced, and AFM resources are concentrated where they are most informative. Although SimuScan does not currently automate sample preparation, laser alignment, or initial feedback tuning, it automates feature search, prioritization, and adaptive scan placement. Depending on the task, automated feature finding and exploration can account for more than 80% of total measurement time, enabling scalable, feedback-driven AFM experimentation with minimal operator input.

2.4. Autonomous AFM of Bacterial Populations

Having validated SimuScan-trained models across diverse systems and within semi-autonomous workflows, we next applied the approach to large-area AFM imaging of bacterial populations. We focused first on *Pseudomonas aeruginosa* PAO1, a rod-shaped Gram-negative bacterium whose well-defined morphology makes it well suited for AFM-based detection and quantitative analysis. Chemically fixed cells were imaged in air and included examples at different stages of the cell cycle, enabling visualization of structural variations associated with elongation, septum formation, and cytokinesis.

In Figure 6, the semi-autonomous AFM pipeline was used to detect and image *P. aeruginosa* cells at subcellular resolution, demonstrating robustness to heterogeneous, soft, and irregular biological features (see Supplementary Figure S13 for detailed single-cell analysis). Using a YOLO-based model trained exclusively on SimuScan data, the system autonomously located, scanned, and segmented 100 cells across a large substrate area (Figure 6a). Morphological variability was quantified by aligning cell

perimeters along the major axis and sorting them by length (Figure 6b). Averaging across the dataset yielded a reference morphology, while height profiles extracted along the longitudinal axis (Figure 6c) revealed systematic trends in size and symmetry. Representative cells (1st, 10th, 20th … 100th by length) illustrate population-level heterogeneity (Figure 6e).

Histograms of major and minor axis lengths (Figure 6d) showed a narrowly distributed cell width (~747 ± 4 nm) and a broader, skewed distribution of cell lengths (~1.3 µm), reflecting different stages of the division cycle. Shorter pre-septation cells exhibited greater height (~200 nm) than elongated, septating cells, which appeared flattened at the division site, consistent with prior AFM and electron microscopy studies of bacterial cytokinesis.[47,48] These morphological signatures indicate that AFM can serve as a quantitative, label-free marker of cell-cycle progression in high-throughput measurements.

Supplementary Figure S13 further reveals fine structural detail, including flagella and cell wall features, and shows that membrane bulges are preferentially localized near the bacterial poles. This recurrent polar localization suggests an intrinsic structural or compositional origin, although local curvature and drying-related effects during sample preparation may also contribute.[49]

This level of population-scale analysis is enabled by the large, systematically acquired datasets produced by the semi-autonomous workflow. By coupling SimuScan-trained models with automated acquisition, the microscope can sample enough objects to support statistically meaningful measurements that are difficult to achieve with conventional, manually guided AFM, thereby extending AFM from more of a single-object characterization toward population-level analysis.

Next, we applied the workflow to a mixed bacterial community containing both bacillus and coccus morphotypes distributed across a >0.3 mm² surface (Figure 7a). The coccus population of *Staphylococcus aureus* NCTC 8325, whose spherical cells commonly occur singly, in chains, or in small clusters, providing a complementary morphology to the rod-shaped *P. aeruginosa*. The resulting high-resolution topography captured hundreds of individual cells with nanoscale contrast. Real-time object detection on tiled image regions using the YOLO classifier was performed with a confidence threshold of 70%, achieving robust detection for both rod-shaped bacillus and spherical coccus cells (Figure 7b). Additional zoomed-in views of representative regions, together with the corresponding detection and segmentation predictions, are shown in Supplementary Figure S14. Spatial mapping of detected centroids (Figure 7c) clearly separated bacillus (red) and coccus (blue) populations, enabling construction of density maps for each morphotype. Bacillus cells were broadly distributed with local clustering (Figure 7d), while coccus cells were sparser but showed pronounced clustering in discrete regions (Figure 7e).

Notably, these results were obtained without extensive task-specific optimization of the SimuScan data generator or model hyperparameters; while further tuning could improve precision for particular samples, the aim here is to demonstrate generalizable performance across heterogeneous microbial systems. Within this autonomous AFM workflow, SimuScan-trained models enable label-free detection and mapping based directly on surface morphology, providing a scalable approach for exploring microbial interfaces with minimal human intervention. In contrast to genetic assays that rely on nucleic acid extraction and amplification, this AFM-based strategy operates directly on physical structure, supporting applications such as surface contamination mapping, pathogen detection, and in situ characterization of microbial populations.

In summary, we introduce SimuScan, a physics-guided synthetic data framework that enables annotation-free training of deep learning models for AFM feature identification, segmentation, and targeted imaging. By generating realistic AFM images with exact ground-truth labels while explicitly modeling instrument-specific artifacts, SimuScan addresses a key barrier to applying machine learning in AFM: the dependence on large, manually labeled experimental datasets.

Across diverse experimental systems, models trained exclusively on SimuScan data generalize reliably to real AFM images, enabling label-free quantification of nanoscale features and supporting workflows that scale from single-object analysis to population-level mapping on heterogeneous surfaces. The ability to define object geometry through either CAD models or self-seeded experimental data allows SimuScan to adapt rapidly to new materials and morphologies without requiring analytic shape models or extensive manual annotation.

Looking forward, SimuScan provides a foundation for integrating adaptive simulation, automated acquisition, and closed-loop decision making in AFM. By coupling domain-specific synthetic data with real-time microscope control, this approach opens a path toward AFM experiments that are more scalable, reproducible, and data-driven, enabling autonomous discovery and prioritization of nanoscale structure across complex biological and materials systems.

3. Methods

SimuScan Data Generation:

Synthetic datasets were generated using SimuScan, a Python (v.3.10)-based package designed to create realistic AFM topographic images by explicitly incorporating typical imaging artifacts. Each synthetic dataset is associated with a dedicated JSON/YAML file that governs image generation. These configuration files can contain more than 500 parameters specifying object properties such as height,

eccentricity, geometric transformations (e.g., bending), and surface texture. In this way, every dataset is fully documented with its statistical parameter distribution, ensuring both reproducibility and traceability.

Beyond dataset generation, SimuScan also provides tools for format conversion and model development, including utilities to transform datasets from COCO annotations into U-Net or YOLO formats, as well as functions for model training and validation. To facilitate broad adoption, SimuScan is distributed as an easy-to-use standalone command-line executable, allowing users to generate synthetic datasets and perform associated processing without requiring a dedicated Python environment. The executable and example configurations are openly available via Zenodo (see Data Availability).

Model Training:

For each framework, two models were trained independently: one dedicated to the detection of regular geometric shapes and another for the detection of bacterial morphologies (bacillus and coccus). The architectures selected were YOLOv8 (Ultralytics, Jocher et al., 2023),[50] a state-of-the-art object detection framework optimized for real-time applications, U-Net[51], a widely adopted semantic segmentation model based on an encoder–decoder architecture with skip connections, and Mask R-CNN[52], an extension of Faster R-CNN that integrates object detection with instance-level segmentation. The synthetic datasets were generated with SimuScan, a Python (v.3.10) - based package and consisted of more than 5,000 images per dataset. Each dataset generation required approximately two hours, and the resulting data were partitioned into training (70%), validation (20%), and test (10%) subsets. All datasets are openly available on Zenodo (DOI: *https://doi.org/10.5281/zenodo.17037230*).

Model training was carried out for 100 epochs on an NVIDIA RTX 2000 Ada GPU (16 GB VRAM, 2.13 GHz), requiring approximately eight hours per model. Typical hyperparameters included a batch size of 8-32, an initial learning rate of 1e-3 with cosine annealing scheduling, and the Adam optimizer with default momentum and weight decay. The metrics obtained on the synthetic test datasets were consistently high, indicating that the training process was successful and that all models converged reliably under the chosen conditions. This setup ensured stable convergence across all models and enabled reproducible benchmarking under consistent computational settings. All models are openly available on Zenodo (DOI: https://doi.org/10.5281/zenodo.17179725).

Nanostructured Surfaces:

A combination of electron beam lithography (EBL) and an etch process were used to fabricate the micro/nanostructured surface. A silicon substrate was cleaned with acetone and IPA prior to spin-coating

it with ma-N 2403 electron beam resist (Micro resist technology GmbH, Germany) at 3000 rpm for 45 s, and baked on a hotplate for 1 minutes at 95° C. Electron beam lithography was performed with a JEOL JBX8100-FS system operating at 100 kV acceleration voltage. A 40 nA beam current and 300 μC/cm^2 dose were used to define the geometry. The patterned substrate was developed in Microposit MF-319 (≤5% tetramethylammonium hydroxide) for 1 min, rinsed with DI water, and dried with nitrogen. The patterned sample was then etched in an inductively couple plasma ion etching system (Oxford Plasmalab 100). The process was carried out in a mixture of 58 sccm C_4F_8, 25 sccm SF_6, and 5 sccm Ar gases at 20°C for 30 s. Removal of the e-beam resist was done by soaking the sample in acetone. Similar fabrication process was carried out for the leaves structures with the resist spun at 2000 rpm, beam current at 2nA and 1 min etching time.

Bacterial Cells and AFM sample:

Pseudomonas aeruginosa PAO1 cells were inoculated into MOPS + glucose minimal medium and incubated overnight at 37 °C with shaking. Cells were then harvested by centrifugation at 7000 *g* for 12 minutes and resuspended in 4% paraformaldehyde in PBS for 15 minutes to fix. Cells were subsequently washed and resuspended in 1X PBS. *Staphylococcus aureus* NCTC 8325 cells were inoculated into LB broth and grown overnight at 37 °C with shaking. Cells were harvested by centrifugation at 7000 *g* for 10 minutes and resuspended in 4% paraformaldehyde in PBS for 15 minutes to fix. Cells were subsequently washed and resuspended in 1X PBS.

Bacterial suspensions were diluted tenfold, and a 20 μL droplet was deposited onto freshly cleaved mica for 10 min. Subsequently, the PBS concentration was gradually reduced every minute (90%, 70%, 50%, 30%, 20%, and 10%) to minimize potential osmotic effects without inducing dehydration. The remaining liquid was gently removed using filter paper, and the samples were dried in a desiccator for 2 h prior to AFM imaging.

DNA preparation and AFM imaging:

Small DNA circles (minicircles) of 339 and 251 bp were prepared using bacteriophage λ-Int site-specific recombination in vivo[53]. Linear forms were obtained via digestion with restriction enzyme NdeI. DNA minicircles were adsorbed onto freshly cleaved mica specimen disks at room temperature, in a buffer containing Ni^{2+} (20 mM HEPES, 3 mM $NiCl_2$, pH 7.4). 2-5 ng of DNA minicircles was added to 10μL of buffer solution and adsorbed for 30 min and then washed four times using the same buffer solution. AFM measurements of DNA minicircles were performed in liquid on Multimode 8 and FastScan Bio AFM

systems in PeakForce Tapping mode, using Peakforce HiResB and FastScan D probes. AFM images were taken at a PeakForce Tapping amplitude of 10 nm with PeakForce frequencies of 4-8 kHz and PeakForce setpoints from 5–20 mV at 512 × 512 pixels to ensure a resolution ≥1 nm/pixel.

Triangular DNA origami (Tilibit Nanosystems Rothemund Triangle, updated design) was folded by mixing the M13 scaffold with staple strands at a 1:10 scaffold:staple molar ratio in folding buffer (5 mM Tris, 1 mM EDTA, 12.5 mM $MgCl_2$, pH 8.0). The mixture was thermally annealed from 80 °C to 25 °C over ~16 h. Folded structures were purified from excess staples using 100 kDa centrifugal filters (Amicon Ultra) with buffer exchange into 10 mM Tris, 10 mM $MgCl_2$ (pH 8.0). Purified origami was diluted to 2 nM in deposition buffer (10 mM Tris, 10 mM $MgCl_2$, pH 8.0) and 10 µL was placed onto freshly cleaved mica for 2–3 min. The surface was gently rinsed with imaging buffer to remove unbound DNA and dried under a gentle flow or nitrogen.

AFM Experiments:

Besides the DNA minicircles, all other AFM images were acquired using the DriveAFM system from Nanosurf, a device equipped with a motorized stage that enables the scanning of large areas (>10 mm^2), ideal for detailed studies of extensive bacterial biofilms. The microscope was controlled via a custom Python-based interface that interacts directly with the microscope's software, allowing users to select the detection model, specify the class of interest, define stopping criteria, and configure imaging parameters such as low- and high-resolution scans, scan speed, and other acquisition settings. For imaging, Multi75-G tips (Budget Sensors) were used in all experiments. Scans were performed in tapping mode with amplitude modulation, maintaining a setpoint of 60%.

Autonomous Workflow:

The autonomous AFM workflow integrates image acquisition, real-time object detection, and guided navigation to new scan areas. AFM imaging was performed on the Nanosurf DriveAFM, equipped with a motorized XY stage and controlled via a custom Python (v3.10) graphical interface that communicates directly with the microscope control software. The interface allows non-expert users to select a pre-trained YOLO object detection model, choose the class of interest to be captured at high resolution, and adjust scanning parameters such as resolution, scan size, and scan speed. Once a region of the sample is selected, the system performs tip approach and acquires a low-resolution survey image, typically 128 × 128 pixels over a 25 × 25 µm field of view. After this survey image is acquired, the system detects objects of the selected class and returns new exploration areas to the interface with updated high-

resolution scan parameters (typically set to 1.5 times larger than the detected object and acquired at 512 × 512 pixels).

Survey and high-resolution images are processed on-the-fly by the YOLO model, which outputs class-specific bounding boxes together with detection confidence scores. Based on these predictions, the workflow automatically selects regions that satisfy predefined criteria. In the case study presented in Figure 4, the selection criterion was defined as a minimum number of isolated bacteria with less than 20% overlap with other objects and a detection confidence >70%, ensuring that only well-isolated cells of sufficient quality were targeted. The control software then repositions the AFM stage to acquire a new low-resolution image and repeats the process, systematically exploring large areas of the sample until a stopping condition is reached.

The system incorporates user-defined stopping rules, such as a maximum number of detected objects, total scanned area, or elapsed acquisition time. Once the stopping condition is satisfied, the workflow terminates autonomously, saving all raw AFM data, detection results, and metadata. This integration of hardware control, real-time inference, and adaptive scan planning enables large-area AFM mapping in a fully unsupervised manner with minimal human intervention.

Data Availability Statement

The datasets and trained models generated in this study have been deposited in the Zenodo repository (https://zenodo.org/records/17037230 and https://zenodo.org/records/-17179726). All source data supporting the figures are provided with this paper.

Code Availability

The analysis and training scripts are available on GitHub at https://github.com/Rmillansol/SimuScan-AFfMtools.git. A citable archived version of the code has been deposited in Zenodo (https://doi.org/10.5281/zenodo.18665434)[54]. The SimuScan synthetic data generator is distributed as a standalone command-line executable and is publicly available via Zenodo (https://zenodo.org/records/18134911), together with example configuration files and documentation.

References

1 Binnig, G., Quate, C. F. & Gerber, C. Atomic force microscope. *Physical review letters* **56**, 930 (1986).

2 Loos, J. The art of SPM: Scanning probe microscopy in materials science. *Advanced Materials* **17**, 1821-1833 (2005).

3 Checa, M., Neumayer, S. M., Tsai, W.-Y. & Collins, L. in *Atomic Force Microscopy for Energy Research* 45-104 (CRC Press, 2022).

4 Dufrêne, Y. F. *et al.* Imaging modes of atomic force microscopy for application in molecular and cell biology. *Nature nanotechnology* **12**, 295-307 (2017).

5 Muzio, M. D., Millan-Solsona, R., Borrell, J. H., Fumagalli, L. & Gomila, G. Cholesterol effect on the specific capacitance of submicrometric DOPC bilayer patches measured by in-liquid scanning dielectric microscopy. *Langmuir* **36**, 12963-12972 (2020).

6 Lozano, H. *et al.* Electrical properties of outer membrane extensions from Shewanella oneidensis MR-1. *Nanoscale* **13**, 18754-18762 (2021).

7 Ahmad Khalili, A. & Ahmad, M. R. A review of cell adhesion studies for biomedical and biological applications. *International journal of molecular sciences* **16**, 18149-18184 (2015).

8 Checa, M., Millan-Solsona, R., Glinkowska Mares, A., Pujals, S. & Gomila, G. Dielectric imaging of fixed HeLa cells by in-liquid scanning dielectric force volume microscopy. *Nanomaterials* **11**, 1402 (2021).

9 Collins, L. *et al.* Perspectives in Scanning Probe Microscopy from the 2021 Joint International Scanning Probe Microscopy and Scanning Probe Microscopy on Soft and Polymeric Materials Conference. *Microscopy and Analysis* **57** (2021).

10 Millan-Solsona, R. *et al.* Analysis of biofilm assembly by large area automated AFM. *npj Biofilms and Microbiomes* **11**, 75 (2025).

11 Kalinin, S. V. *et al.* Automated and autonomous experiments in electron and scanning probe microscopy. *ACS nano* **15**, 12604-12627 (2021).

12 Sadeghian, H. *et al.* Automated cantilever exchange and optical alignment for high-throughput parallel atomic force microscopy. *IEEE/ASME Transactions on Mechatronics* **22**, 2654-2661 (2017).

13 Yoo, R. Y. The Story behind the First Automatic Atomic Force Microscope. *Microscopy Today* **30**, 40-45 (2022).

14 Casuso, I. & Scheuring, S. Automated setpoint adjustment for biological contact mode atomic force microscopyimaging. *Nanotechnology* **21**, 035104 (2009).

15 Liu, H. *et al.* Intelligent tuning method of PID parameters based on iterative learning control for atomic force microscopy. *Micron* **104**, 26-36 (2018).

16 Abramovitch, D. Y., Hoen, S. & Workman, R. in *2008 American Control Conference.* 2684-2689 (IEEE).

17 Zhou, X., Dong, X., Zhang, Y. & Fang, Y. in *2009 IEEE Control Applications,(CCA) & Intelligent Control,(ISIC).* 1271-1275 (IEEE).

18 Millan-Solsona, R. *et al.* Analysis of biofilm assembly by large area automated AFM. *npj Biofilms and Microbiomes* **11**, 75, doi:10.1038/s41522-025-00704-y (2025).

19 Liu, Y. *et al.* AEcroscopy: A Software–Hardware Framework Empowering Microscopy Toward Automated and Autonomous Experimentation. *Small Methods* **8**, 2301740, doi:https://doi.org/10.1002/smtd.202301740 (2024).

20 Checa, M. *et al.* High-speed mapping of surface charge dynamics using sparse scanning Kelvin probe force microscopy. *Nature Communications* **14**, 7196 (2023).

21 Degenhardt, J., Bounaim, M. W., Deng, N., Tutsch, R. & Dai, G. A new kind of atomic force microscopy scan control enabled by artificial intelligence: Concept for achieving tip and sample safety through asymmetric control. *Nanomanufacturing And Metrology* **7**, 11 (2024).

22 Wei, Z. *et al.* PICTS: A Novel Deep Reinforcement Learning Approach for Dynamic PI Control in Scanning Probe Microscopy. *arXiv preprint arXiv:2502.07326* (2025).

23 Liu, Y. *et al.* Machine Learning-Based Reward-Driven Tuning of Scanning Probe Microscopy: Toward Fully Automated Microscopy. *ACS nano* **19**, 19659-19669 (2025).

24 Liu, Y. & Kalinin, S. V. Pareto-Optimal Experimentation: Human-Guided Multi-Objective Bayesian Optimization in Scanning Probe Microscopy. *Nano Letters* (2025).

25 Paruchuri, A., Wang, Y., Gu, X. & Jayaraman, A. Machine learning for analyzing atomic force microscopy (AFM) images generated from polymer blends. *Digital Discovery* **3**, 2533-2550 (2024).

26 Yablon, D. *et al.* Deep learning to establish structure property relationships of impact copolymers from AFM phase images. *Mrs Communications* **11**, 962-968 (2021).

27 Checa, M., Millan‐Solsona, R., Mares, A. G., Pujals, S. & Gomila, G. Fast label‐free nanoscale composition mapping of eukaryotic cells via scanning dielectric force volume microscopy and machine learning. *Small Methods* **5**, 2100279 (2021).

28 Vasudevan, R. K. *et al.* Autonomous experiments in scanning probe microscopy and spectroscopy: choosing where to explore polarization dynamics in ferroelectrics. *ACS nano* **15**, 11253-11262 (2021).

29 Rade, J. *et al.* Deep learning for live cell shape detection and automated afm navigation. *Bioengineering* **9**, 522 (2022).

30 Sotres, J., Boyd, H. & Gonzalez-Martinez, J. F. Enabling autonomous scanning probe microscopy imaging of single molecules with deep learning. *Nanoscale* **13**, 9193-9203 (2021).

31 Huang, B., Li, Z. & Li, J. An artificial intelligence atomic force microscope enabled by machine learning. *Nanoscale* **10**, 21320-21326 (2018).

32 Deng, J. *et al.* in *2009 IEEE conference on computer vision and pattern recognition.* 248-255 (Ieee).

33 Kirillov, A. *et al.* in *Proceedings of the IEEE/CVF international conference on computer vision.* 4015-4026.

34 Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data* **6**, 1-48 (2019).

35 Nečas, D. & Klapetek, P. Synthetic data in quantitative scanning probe microscopy. *Nanomaterials* **11**, 1746 (2021).

36 Klapetek, P. & Ohlídal, I. Theoretical analysis of the atomic force microscopy characterization of columnar thin films. *Ultramicroscopy* **94**, 19-29 (2003).

37 Klapetek, P., Ohlídal, I. & Bílek, J. Influence of the atomic force microscope tip on the multifractal analysis of rough surfaces. *Ultramicroscopy* **102**, 51-59 (2004).

38 Klapetek, P., Valtr, M., Nečas, D., Salyk, O. & Dzik, P. Atomic force microscopy analysis of nanoparticles in non-ideal conditions. *Nanoscale research letters* **6**, 514 (2011).

39 Chen, Y. Improving dimensional measurement from noisy atomic force microscopy images by non‐local means filtering. *Scanning* **38**, 113-120 (2016).

40 Wang, Y.-F. *et al.* Double-tip artifact removal from atomic force microscopy images. *IEEE Transactions on Image Processing* **25**, 2774-2788 (2016).

41 Argento, C. & French, R. Parametric tip model and force–distance relation for Hamaker constant determination from atomic force microscopy. *Journal of Applied Physics* **80**, 6081-6090 (1996).

42 Niina, T., Fuchigami, S. & Takada, S. Flexible fitting of biomolecular structures to atomic force microscopy images via biased molecular simulations. *Journal of Chemical Theory and Computation* **16**, 1349-1358 (2020).

43 Fuchigami, S., Niina, T. & Takada, S. Case report: Bayesian statistical inference of experimental parameters via biomolecular simulations: atomic force microscopy. *Frontiers in Molecular Biosciences* **8**, 636940 (2021).

44 Alldritt, B. *et al.* Automated structure discovery in atomic force microscopy. *Science advances* **6**, eaay6913 (2020).
45 González Lastre, M. *et al.* Molecular identification via molecular fingerprint extraction from atomic force microscopy images. *Journal of Cheminformatics* **16**, 130 (2024).
46 Beton, J. G. *et al.* TopoStats–A program for automated tracing of biomolecules from AFM images. *Methods* **193**, 68-79 (2021).
47 Navarro, P. P. *et al.* Cell wall synthesis and remodelling dynamics determine division site architecture and cell shape in Escherichia coli. *Nature Microbiology* **7**, 1621-1634 (2022).
48 Van Der Hofstadt, M. *et al.* Internal hydration properties of single bacterial endospores probed by electrostatic force microscopy. *ACS nano* **10**, 11327-11336 (2016).
49 Lithgow, T., Stubenrauch, C. J. & Stumpf, M. P. Surveying membrane landscapes: a new look at the bacterial cell surface. *Nature Reviews Microbiology* **21**, 502-518 (2023).
50 Ultralytics YOLOv8, https://github.com/ultralytics/ultralytics (2023).
51 Ronneberger, O., Fischer, P. & Brox, T. in *International Conference on Medical image computing and computer-assisted intervention.* 234-241 (Springer).
52 He, K., Gkioxari, G., Dollár, P. & Girshick, R. in *Proceedings of the IEEE international conference on computer vision.* 2961-2969.
53 Pyne, A. L. *et al.* Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides. *Nature Communications* **12**, 1053 (2021).
54 Millan-Solsona, R. SimuScan-AFMtools (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.18665434 (2026).

Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science FWP ERKCZ64, Structure Guided Design of Materials to Optimize the Abiotic-Biotic Material Interface, as part of the the Biopreparedness Research Virtual Environment (BRaVE) initiative. AFM measurements, sample preparation, and image analysis were conducted as part of a user project at the Center for Nanophase Materials Sciences (CNMS), which is a US Department of Energy, Office of Science User Facility at Oak Ridge National Laboratory. We also acknowledge support from a UKRI Future Leaders Fellowship (MR/W00738X/1) and the Henry Royce Institute for Advanced Materials, funded through EPSRC grants EP/R00661X/1, EP/S019367/1, EP/P02470X/1 and EP/P025285/1 (A.L.B.P.).

Author contributions

R.M.-S. and L.C. conceived the study (Conceptualization) and developed the methodology. R.M.-S. led the software development, while validation and formal analysis was assisted by M.C., S.S.M., R.K.V., and L.C. Experimental investigation was performed by R.M.-S.. Resources were provided by S.R.B., B.S., A.N.B., J.L.M.-F., A.P., and L.W. Data curation was handled by R.M.-S., M.C., and L.C. Data visualization was conducted by R.M.-S., M.C., and L.C.. L.C. supervised and administered the

project. The original draft was written by L.C. and R.M.-S., and all authors reviewed and edited the manuscript. Funding was acquired by L.C. and S.R.

Competing interests

The authors declare no competing interests.

Table 1. Performance of models trained on synthetic datasets generated with SimuScan. All reported metrics were computed by directly comparing model predictions on experimental AFM images against manually annotated ground-truth object instances and pixel masks, which were used exclusively for evaluation and not for model training. Feature-level performance is quantified using AP@0.5 for object detection and intersection-over-union (IoU) for pixel-level segmentation. Global object-level precision, recall, and F1-score summarize overall detection reliability and missed-object rates. Metrics are reported separately for nanostructures and bacterial morphologies.

Category	Evaluation Level	Class / Metric	YOLOv8	U-Net	Mask R-CNN
Nanostructures	**Feature**	Background (IoU)	0.99	0.99	0.99
		Rectangle (AP@0.5 / IoU)	0.97 / 0.85	— / 0.84	0.68 / 0.61
		Hexagon (AP@0.5 / IoU)	0.99 / 0.84	— / 0.86	0.95 / 0.86
		Circle (AP@0.5 / IoU)	0.98 / 0.90	— / 0.88	0.96 / 0.85
	Global	Precision	0.91	0.89	0.98
		Recall	0.98	0.97	0.81
		F1-score	0.94	0.93	0.89
Bacteria	**Feature**	Background (IoU)	0.99	0.99	0.99
		Bacillus (AP@0.5 / IoU)	0.82 / 0.80	— / 0.79	0.66 / 0.57
		Coccus (AP@0.5 / IoU)	0.91 / 0.71	— / 0.73	0.52 / 0.71
	Global	Precision	0.81	0.86	0.80
		Recall	0.93	0.88	0.79
		F1-score	0.87	0.87	0.80

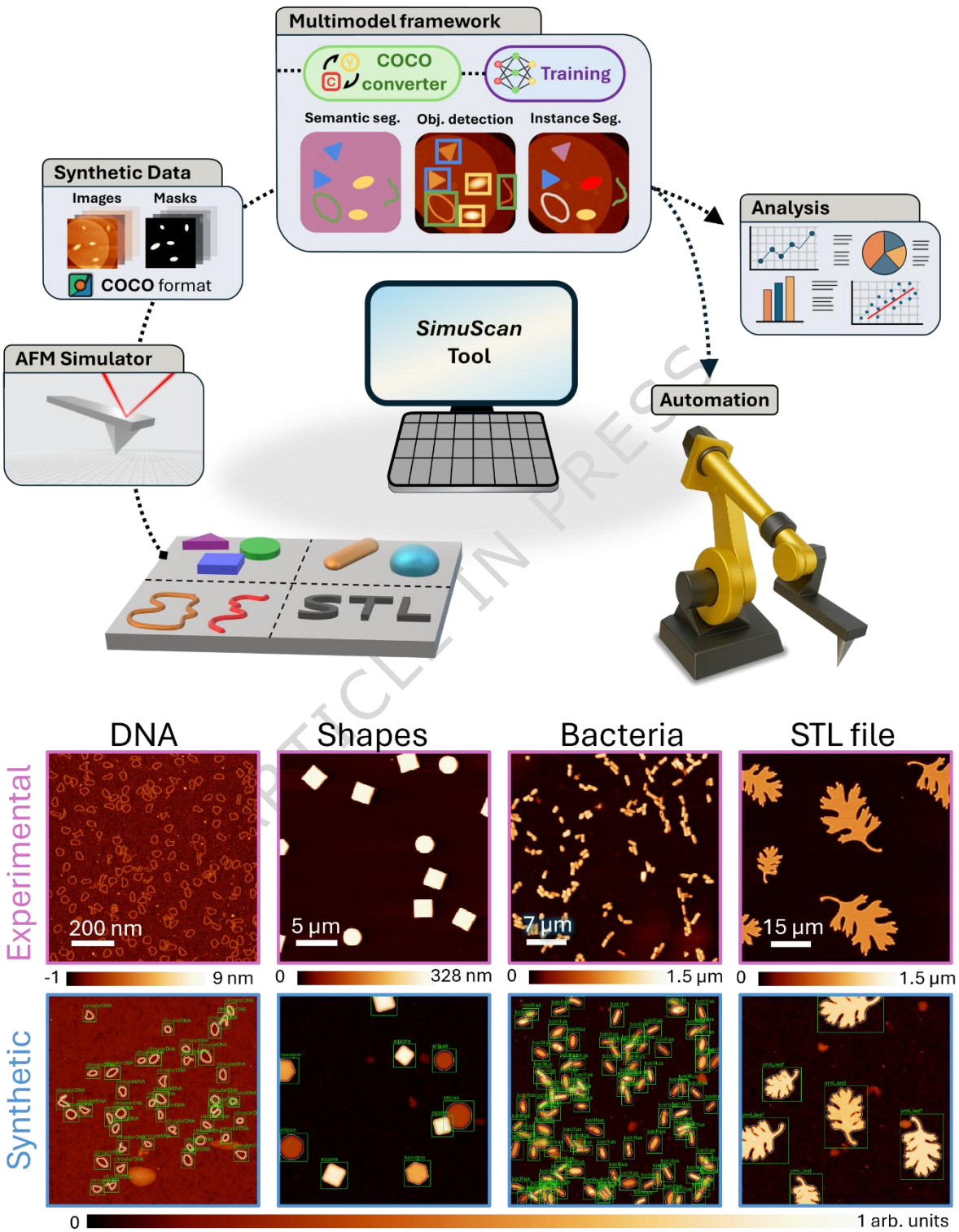

Figure 1. Semi-autonomous AFM workflow enabled by SimuScan. SimuScan integrates synthetic AFM data generation with deep-learning segmentation models, including semantic, object detection, and

instance segmentation, to enable semi-autonomous AFM operation. Synthetic topographies derived from defined shapes or arbitrary STL geometries are converted into labeled datasets, used to train segmentation models, and deployed for real-time feature recognition and automated microscope control. Bottom panels show experimental and corresponding synthetic AFM images of DNA assemblies, nanostructures, bacterial cells, and leaf-shaped features, spanning length scales from the nanoscale to the microscale. Synthetic images use normalized height values (0–1) and pixel-based coordinates, with experimental data rescaled to match prior to segmentation.

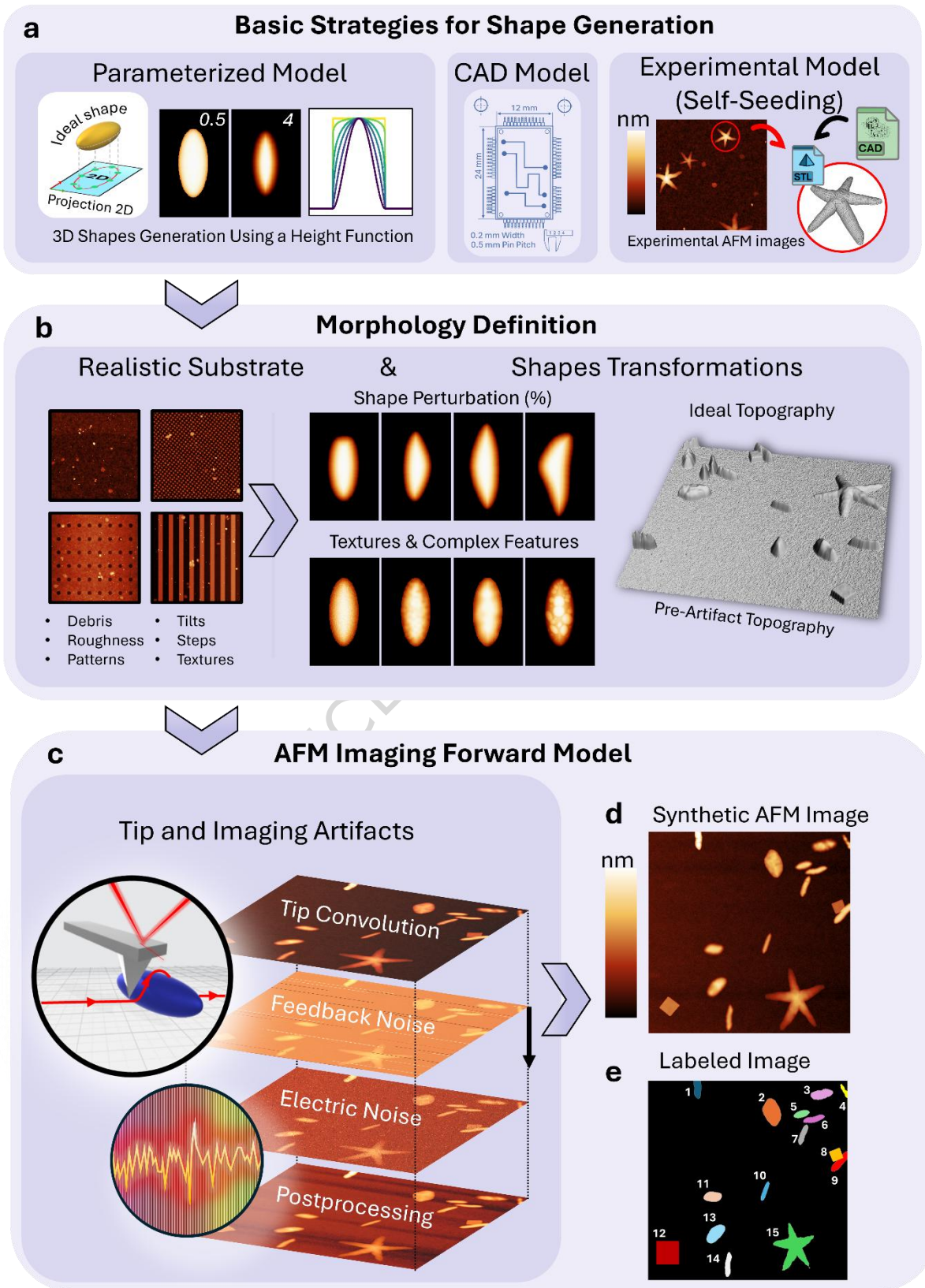

Figure 2. SimuScan synthetic AFM data generation pipeline. **(**a) Basic strategies for shape generation, including parameterized object models, CAD-derived geometries (representative schematic example

shown), and self-seeded experimental features extracted directly from AFM images and converted to STL objects. (b) Morphological definition through realistic substrate generation and shape transformations, including roughness, steps, tilts, textures, debris, and controlled shape perturbations, producing an ideal pre-artifact topography. Further details on the ideal parameterized model formulation are provided in Supplementary Figure S1. (c) AFM imaging forward model incorporating tip convolution, feedback and line-flattening artifacts, electrical noise, and post-processing effects to simulate experimental AFM acquisition. (d) Resulting synthetic AFM image after application of the forward model. (e) Corresponding labeled instance map generated directly from the known object geometries, providing pixel-accurate ground truth for training segmentation models.

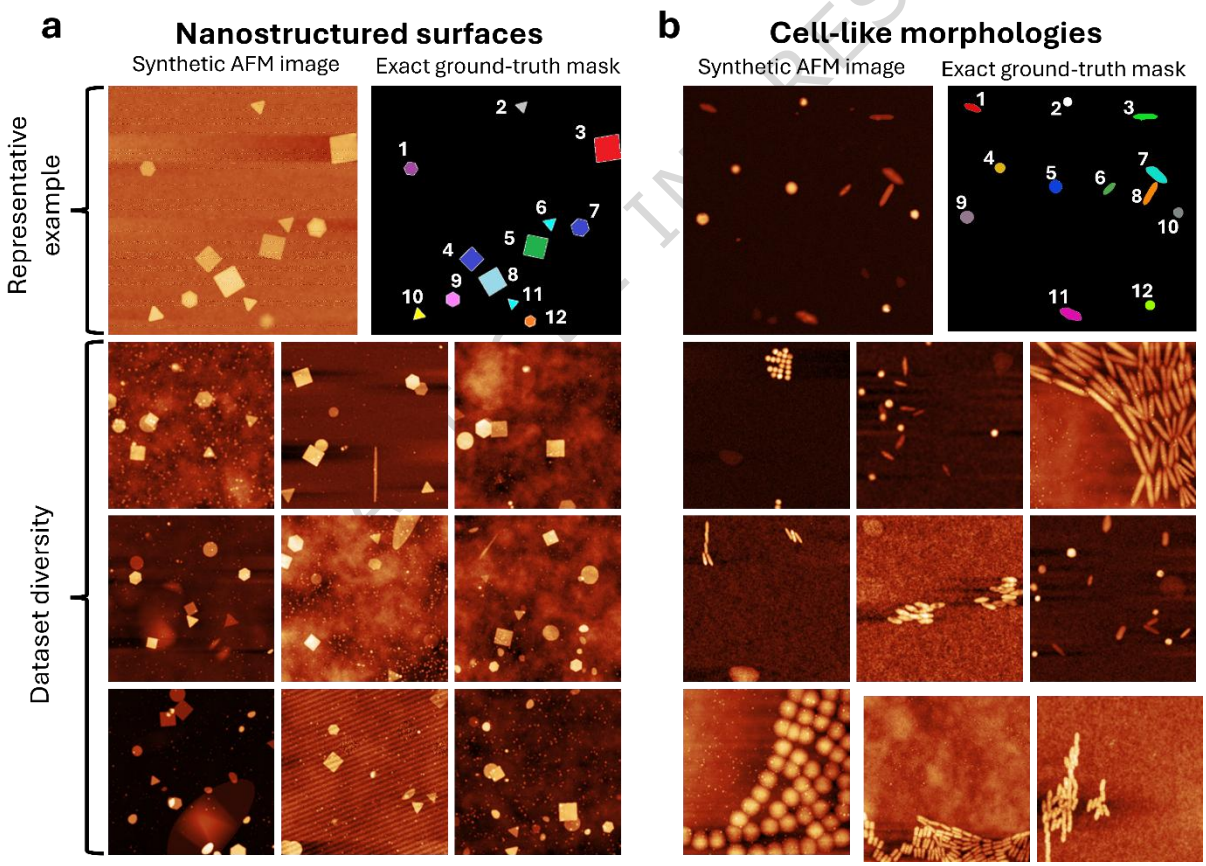

Figure 3. **Examples of SimuScan synthetic AFM datasets**. Example synthetic AFM image (a) Representative synthetic AFM image of a nanostructured surface containing randomly oriented geometric objects, including squares, rods, and irregular shapes, rendered with realistic noise, background structure, and line-flattening artifacts. Corresponding instance labels are shown for each object. (b) Additional synthetic datasets illustrating cell-like and bacterial morphologies, including both randomly distributed

features and clustered colony-like organizations, demonstrating SimuScan's ability to generate diverse and complex nanoscale structures for training segmentation models.

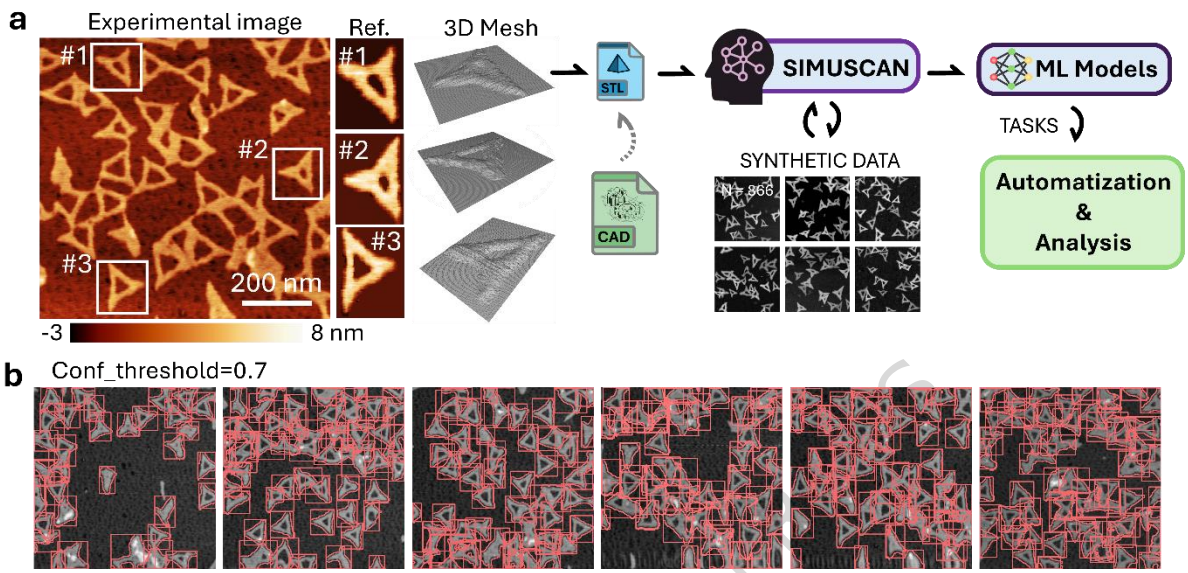

Figure 4. STL-driven synthetic data generation and automated analysis of DNA origami. (a) Schematic illustration of the SimuScan workflow for complex biomolecular structures. Starting from an experimental AFM image of triangular DNA origami, selected topographic regions are extracted and processed to remove background contributions, yielding clean reference geometries. These references are converted into STL files and used to generate realistic synthetic AFM datasets that capture the morphology and imaging artifacts of the experimental system. **(b)** Representative predictions on experimental AFM images of triangular DNA origami using a YOLOv8 model trained exclusively on SimuScan-generated data. Both object detection and instance segmentation accurately identify origami structures, demonstrating robust generalization to experimental data and performance consistent with the quantitative trends reported in the main text.

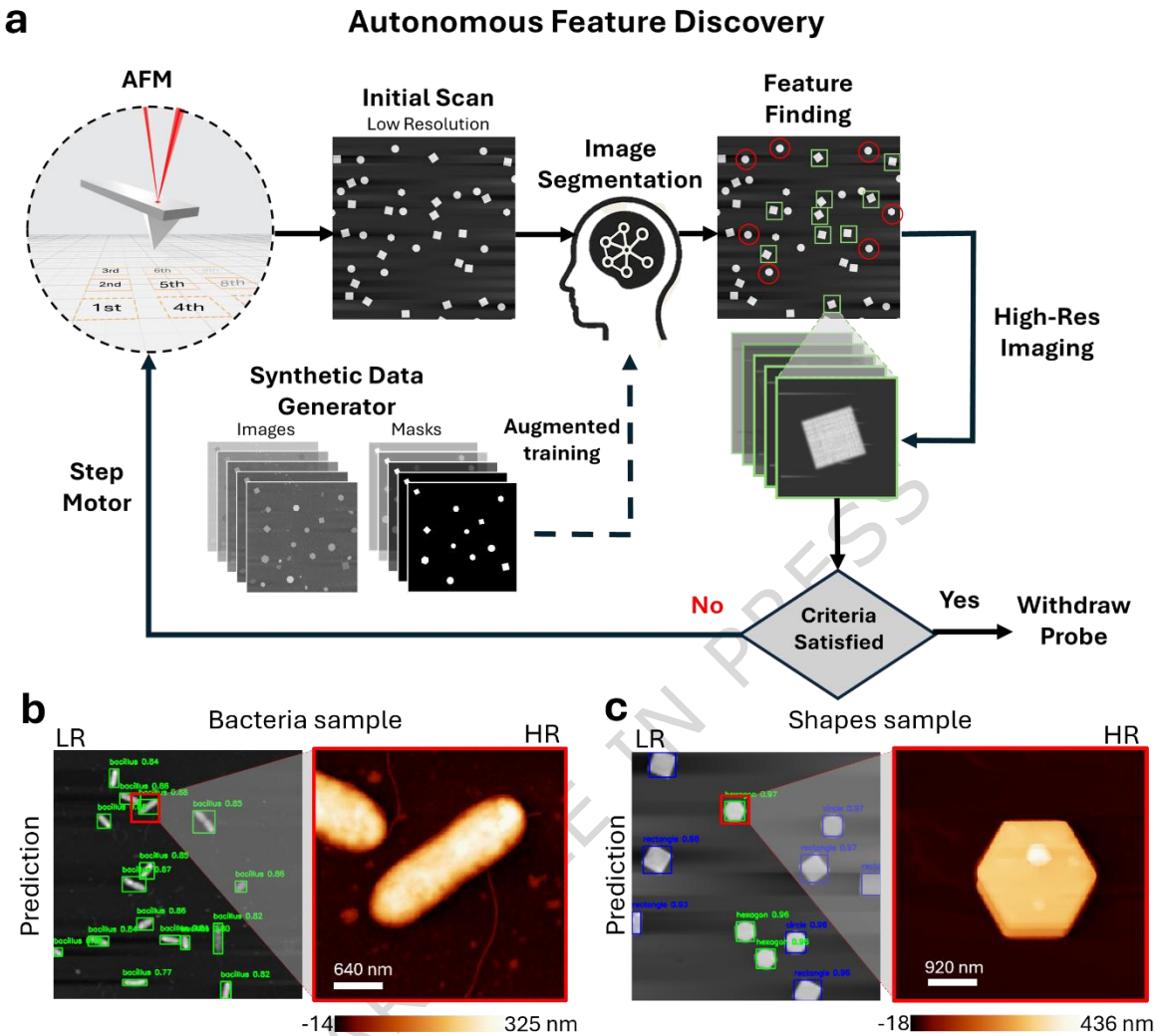

Figure 5. Autonomous feature discovery workflow using a benchmark lithographic sample. (a) Schematic of the closed-loop AFM workflow for autonomous feature discovery. The system begins with a low-resolution (LR) overview scan of a nanostructured silicon test sample patterned with randomly distributed geometric shapes (squares, triangles, circles). Real-time segmentation, trained on synthetic AFM data, identifies and classifies individual features. A feature selection algorithm then prioritizes candidates for high-resolution (HR) imaging. Once user-defined criteria are satisfied (e.g., sufficient number or diversity of features imaged), the probe is withdrawn and the system moves to a new location via automated stage control. This pipeline enables adaptive, high-throughput AFM characterization without manual oversight. In (b) shows an example of a low-resolution and a high-resolution image of bacteria, in (c) shows another example of a sample of nanostructures. (b,c) Representative examples of low-resolution overview scans and corresponding high-resolution acquisitions. Predicted instances are overlaid with bounding markers, where green indicates the target object class considered in the study and

blue indicates additional classes detected by the model but not analyzed here. Labels display the predicted class and confidence score. For clarity, annotation markers are illustrative and not intended to be read in full detail.

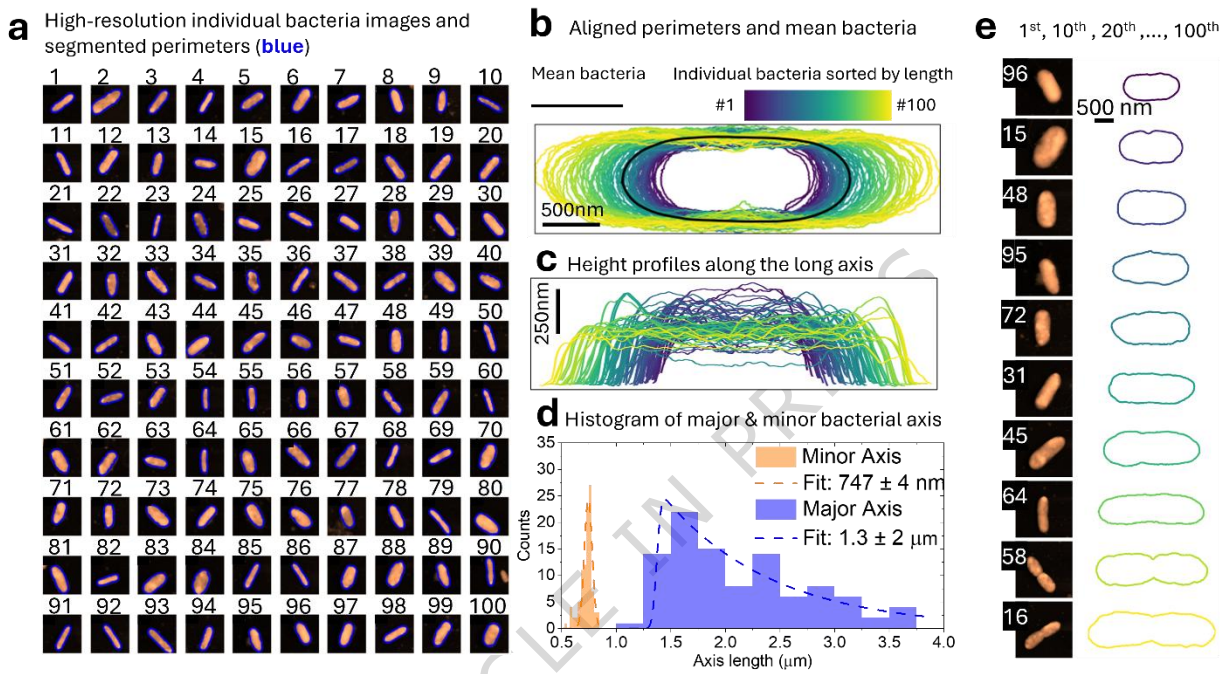

Figure 6. Autonomous high-resolution imaging and morphological analysis of individual Bacillus cells. (a) Dataset of 100 *P. aeruginosa* cells imaged using high-resolution microscopy, with extracted cell perimeters overlaid in blue. Images were autonomously captured and segmented. (b) Cell perimeters, aligned by their long axis and sorted by cell length, reveal structural variation and a mean bacterium shape (black contour). (c) Height profiles extracted along the longitudinal axis of each bacterium demonstrate morphological diversity and are visualized as 2.5D height maps. (d) Histogram of major (blue) and minor (orange) axis lengths, with fitted distributions (SkewNorm for major, Gaussian for minor) capturing population-level variation in cell shape. (e) Representative individual cells (1st, 10th, 20th, …, 100th) and their corresponding perimeters, illustrating morphological heterogeneity across the dataset.

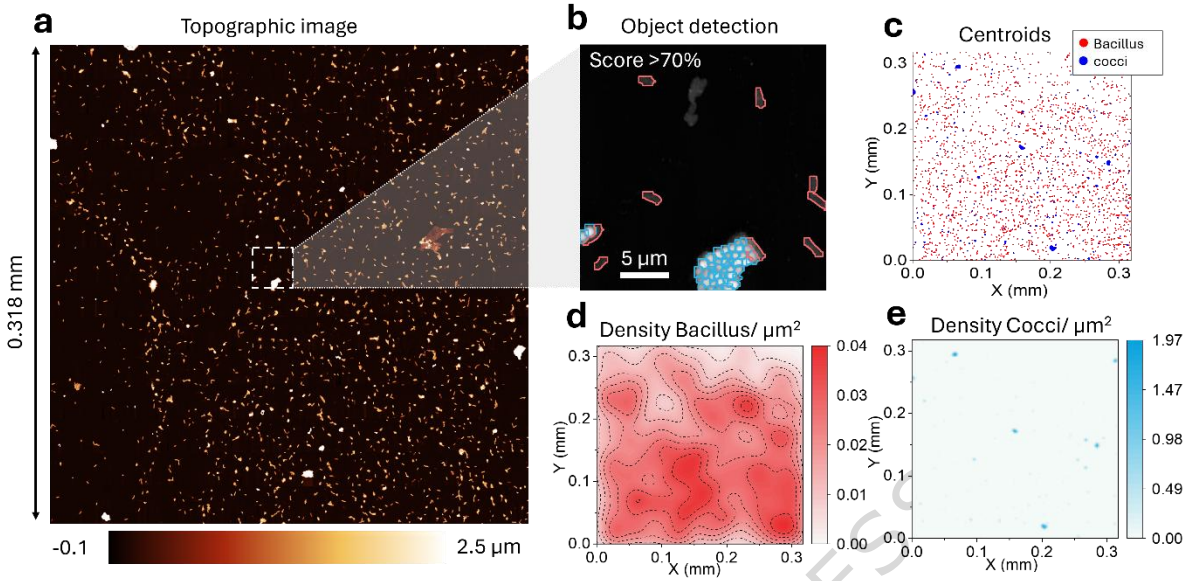

Figure 7. Large-area autonomous mapping and classification of bacterial morphotypes from topographic data. (a) High-resolution topographic image (0.318 mm × 0.318 mm) capturing a mixed bacterial population across a complex surface. (b) Object detection using a trained model with confidence scores >70%, identifying bacilli (outlined in red) and cocci (cyan) morphologies within the inset region. (c) Spatial distribution of detected bacterial centroids over the entire image area, color-coded by class (red: bacilli, blue: cocci). (d) Population density maps of bacillus spatial density, highlighting regions of increased cell concentration and local clustering behavior. (e) Corresponding density map for coccus organisms, indicating their sparse highly clustered presence in this field of view.

Editor's Summary

This study introduces SimuScan, a synthetic data-driven deep learning framework that enables autonomous atomic force microscopy without manual annotation, accelerating AI-based nanoscale characterization.

Peer Review Information: *Nature Communications* thanks Philippe Leclere and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.